Contingency tables

We just got done learning about the goodness of fit test. This test is used for categorical data where we have only one factor. For example, we are looking at blood type, or color or similar. But we often have more than one factor. This is where contingency tables are used.

For example, perhaps we're trying to the distribution of blood types is different in men and women. Now we have *two* factors: (1) blood type, and (2) sex. For blood type we have four categories (A, B, AB, O), and for sex we have two categories (male, female). If we went out and recorded blood types of 153 men and 162 women, we could analyze our data with a χ^2 contingency table test to see if the distribution of blood types is different.

Let's provide a more realistic example. In 1988 the state of Florida collected the following information on the use of seat belts and their relationship to fatalities:

Safety equipment in use	Injury		
	Fatal	Non-fatal	**Total**
None	1,601	165,527	**167,128**
Seat belt	510	412,368	**412,878**
Total	**2,111**	**577,895**	**580,006**

We have two factors to consider, *Injury* and *Safety equipment in use*. Each has two categories or levels. Now that we have some data, what might we be interested in? The obvious question to ask is "do seatbelts save lives?".

We approach this problem by figuring out the *proportion* of people who died while not wearing a seatbelt. We divide the number of people who died while not wearing a seatbelt (1,601) by the *total* number of people not wearing a seatbelt (167,128):

$$\text{Proportion of people who died while not wearing a seatbelt} = \hat{p}_1 = \frac{1,601}{167,128} = 0.00958$$

This doesn't sound so bad. If you're in an accident and you're not wearing a seatbelt you "only" have a 0.96% chance of dying?

Well, we're not done yet. We want to compare this to the proportion of people who died while wearing a seatbelt:

$$\text{Proportion of people who died while wearing a seatbelt} = \hat{p}_2 = \frac{510}{412,878} = 0.00124$$

This proportion is much lower than the other one, and we begin to see that maybe seatbelts are in cars for a reason. But how do we turn this into a hypothesis test? What are we interested in testing? One possibility is to see if the (population) proportion of our two groups are different. Specifically, let's do:

$H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$ (does this alternative really make sense?)

Remember that the sample proportion, $\hat{p}$, estimates p, the true population proportion, so we have $\hat{p}_1$ estimating p_1, and $\hat{p}_2$ estimating p_2.

What about our alternative hypothesis? Are we really interested in $p_1 \neq p_2$? No, we're interested in seatbelts *saving* lives, so really we need a one sided alternative hypothesis. The proportion of fatalities should be *lower* if wearing a seatbelt and we should be using:

$H_1 : p_1 > p_2$

So one way to do a contingency table test is to compare proportions. However, sometimes we are more interested in establishing dependence or independence between our factors.

Suppose we are interested in two species of mice. Does the presence of species A influence the presence of species B? In other words, does species B care if species A is around? Suppose we sample 179 plots along Skyline drive to determine which mice are in each plot. We get the following:

		Species B		
		Present	Absent	**Total**
Species	Present	38	53	**91**
A	Absent	68	20	**88**
	Total	**106**	**73**	**179**

Maybe A and B are in competition, or maybe they actually like each other. What kind of question are we interested in asking? Does it make sense to compare the proportion of species A that is present when species B is present with the proportion of species A that is present when species B is absent? Does the previous sentence sound a bit confusing? That might be an indication that we need to ask the question differently. How about we jump right to our hypotheses and do:

H_0 : Species A and B are independent of each other.

H_1 : Species A and B are *not* independent of each other (they're dependent).

That sounds much easier. All we're asking here is if it makes a different to species A if species B is there, or vice versa.

The important thing to realize is that we have two contingency tables above; the first asks a hypothesis about proportions, the second about independence. It takes a little practice to determine which hypotheses (proportions or independence) you want to use for a particular problem, but the good news is that the analysis for both cases is exactly the same. Let's outline things:

1. Write down your hypotheses:

   $H_0 : p_1 = p_2$

   $H_1 : p_1 \neq p_2$ (or one sided ($<$ or $>$))

   or

   H_0 : Factors A and B are independent of each other.

   H_1 : Factors A and B are *not* independent of each other (they're dependent).

2. Pick a value for α.

3. Calculate your test statistic. This will be the same as for the goodness of fit test, although the value for c is subtly different:

   $$\chi^{2^*} = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i}$$

   where c is now *number of cells*, not categories. Both our tables have four cells (the rows/columns with totals do not count).

   One thing you might wonder is where we get our expected values - more on that below.

4. Finally you compare your value of χ^{2^*} to the χ^2_{table} value using $d.f. = (r-1) \times (k-1)$, where:

   $r =$number of rows (not including totals).

   $k =$number of columns (not including totals).

   So we have: $\chi^2_{table} = \chi^2_{\alpha,(r-1)\times(k-1)}$

5. Then as usual, if $\chi^{2^*} \geq \chi^2_{table}$ (or if $p-$value $\leq \alpha$) we reject H_0.

So how do we get our expected values? Unlike the goodness of fit test, there are no numbers (proportions) in the null hypothesis. We need to calculate our expected values.

Let's use our mouse example (species A versus species B) as an example. We will assume, as usual, that our null hypothesis (H_0) is true. This says that species A shouldn't be affected by species B.

Let's figure out the proportion of plots with species A. To do this, we can just use the totals column. We have:

$$\text{Proportion of plots with species A} = \frac{\text{number of plots with A}}{\text{total number of plots}} = \frac{91}{179} = 0.508$$

In other words 50.8% of plots have species A present.

The null hypothesis implies that it shouldn't make any difference if species B is present or not. So, for example, if we have 106 plots (which is the number of plots with species B present), 50.8% of them should have species A present. That's an easy calculation:

$$0.508 \times 106 = 53.9$$

In other words, if we have 106 plots, we *expect* 58.8 of them to have species A. We just calculated our first expected value and we can add this to the table as follows (in parenthesis and italics):

		Species B		
		Present	Absent	**Total**
Species	Present	38 (*58.9*)	53	**91**
A	Absent	68	20	**88**
	Total	**106**	**73**	**179**

We can do the same for the 73 plots that do not have species B:

$$0.508 \times 73 = 37.1$$

So we *expect* 37.1 plots to have species A when species B is absent, and again we can fill this into our table:

		Species B		
		Present	Absent	**Total**
Species	Present	38 (*58.9*)	53 (*37.1*)	**91**
A	Absent	68	20	**88**
	Total	**106**	**73**	**179**

We can continue the same way for the second row, but let's think a little more about what we're doing. We're taking the row total and dividing this by the grand total and then multiplying this number by the column total. In other words, we are doing:

$$\text{expected value} = \frac{\text{row total}}{\text{grand total}} \times \text{column total}$$

And this gives us a formula that is easy to remember:

$$\text{expected value} = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

So for any expected value, we can simply that the row total for the cell we want, multiply by the column total for the cell we want, and divide this by the grand total:

$$\text{expected value for cell in first row and first column} =$$

$$\frac{\text{row one total} \times \text{column one total}}{\text{grand total}} = \frac{91 \times 106}{179} = 58.9$$

So now we know how to calculate our expected values. Let's do a few examples, starting with the seatbelt/fatality problem:

> H_0: The proportion of people killed is the same whether or not they are wearing a seatbelt.
>
> H_1: The proportion of people killed while not wearing a seatbelt is higher than the proportion killed while wearing a seatbelt.
>
> or, in symbols:
>
> H_0: $p_1 = p_2$
>
> H_1: $p_1 > p_2$
> (See above for the explanation for using a one sided alternative).

Let's pick $\alpha = 0.05$.

Before we move on, we need to make sure that our data agree with our alternative hypothesis (we're doing a one sided test!):

$p_1 = 0.00958 > p_2 = 0.00124$, so yes, our data agree with H_1.

Now we calculate our expected values:

Our first expected values (for row "None" and column "Fatal"):

$$\frac{2,111 \times 167,128}{580,006} = 608.28$$

Our second expected value ("None" and "Non-fatal"):

$$\frac{577,895 \times 167,128}{580,006} = 166,519.72$$

Our third expected value ("Seat belt" and "Fatal"):

$$\frac{2,111 \times 421,878}{580,006} = 1,502.72$$

And finally, our last expected value ("Seat belt" and "Non-fatal"):

$$\frac{577,895 \times 421,878}{580,006} = 411,375.28$$

And we can add these into our table so we have everything in one place:

Safety equipment in use	Injury		
	Fatal	Non-fatal	**Total**
None	1,601 *(608.28)*	165,527 *(166,519.72)*	**167,128**
Seat belt	510 *(1,502.72)*	412,368 *(411,375.28)*	**412,878**
Total	**2,111**	**577,895**	**580,006**

Now we can calculate our value for χ^{2^*}:

$$\chi^{2^*} = \frac{(1,601 - 608.28)^2}{608.28} + \frac{(165,527 - 166,519.72)^2}{166,519.72}$$

$$+ \frac{(510 - 1,502.72)^2}{1,502.72} + \frac{(412,368 - 411,375.28)^2}{411,375.28}$$

$$= 2,284.25$$

Finally, we look up our critical value of χ^2 in our tables and get $\chi^2_{0.05,1} = 2.71$ (make sure you use the one sided value).

From this we conclude that since our critical value is larger than the table value we reject our H_0 and conclude that seatbelts save lives.

Incidentally, R says that our p-value is tiny (we have an absurdly large value of χ^{2^*}): $p < 1.1\text{e} - 16$, so if we compare this to α we reject as expected. (The p-value is so small, R can only give us a maximum value for p).

How about our other example? Let's take a look at our two species of mice, but we'll skip some of the details this time:

H_0 : Species A and B are independent of each other.

H_1 : Species A and B are *not* independent of each other (they're dependent).

Set $\alpha = 0.05$.

Filling in the rest of our expected values in our table:

		Species B		
		Present	Absent	**Total**
Species A	Present	38 (*58.9*)	53 (*37.1*)	**91**
	Absent	68 (*52.1*)	20 (*35.9*)	**88**
	Total	**106**	**73**	**179**

And our χ^{2^*} is (we didn't write down the middle two terms):

$$\chi^{2^*} = \frac{(38 - 58.9)^2}{58.9} + ... + \frac{(20 - 35.9)^2}{35.9} = 23.4$$

And since our value of χ^{2^*} is larger than the (two sided) table value ($\chi^2_{0.05,1} = 3.84$) we reject H_0.

Since we rejected, we want to take this example a step further. Our question now becomes, do the two mice attract each other (are there more of species A present when B is present) or do the repel each other (are there less of species A present when B is present). To do this we need to figure out some proportions. Let's calculate the proportion of species A that is present when species B is present:

$$\hat{p}_1 = \frac{38}{106} = 0.358 \text{ or } 35.8\%$$

Now let's do the proportion of species A that is present when species B is absent:

$$\hat{p}_2 = \frac{53}{73} = 0.726 \text{ or } 72.6\%$$

Which of these is higher? There are more of species A present when species B is absent, so they repel each other (they don't like to share plots).

You may need to calculate proportions like this to figure out which way a relationship in a contingency table is going. The calculation isn't difficult, although sometimes it's a bit confusing to know which numbers to use. It turns out, it doesn't make any difference.

We used the row 1 cell numbers and divided by the column totals. You can use the row 2 cell numbers and divide by the column totals. You can also use the column 1 cell numbers and divide by row totals, or the column 2 cell numbers and divide by row totals.

In other words, you will always get the same answer about attraction or repelling. You may need to think a bit about what the $\hat{p}$'s that you're calculating represent, but as long as you do this, you'll be fine.

What about bigger tables? So far we've looked at table with 2 rows and 2 columns. Bigger tables are referred to ash $R \times K$ tables. This is pretty standard language in statistics, so you'll have to remember that anything bigger than 2×2 is $R \times K$.

The good news is that the math and everything else is the same as for a 2×2 table. Let's do a quick example.

We want to know if there's a difference in food preference in three species of squirrel. We put out peanuts and walnuts for our squirrels and get the following results:

		Nut type		
		Peanut	Walnut	**Total**
Species	A	21	32	**53**
	B	10	15	**25**
	C	15	12	**27**
	Total	**46**	**59**	**105**

H_0 : The proportions of peanuts and walnuts are the same for all three species.

H_1 : The proportions are not the same.

(Note that we can't do a one sided test since our table is bigger than 2×2).

Set $\alpha = 0.05$.

We will not go through all the calculations, but here are a few steps:

> Our first expected value ("Species A" and "Peanut") is $\frac{46 \times 53}{105} = 23.2$
>
> And so on for the other expected values.
>
> Our $\chi^{2^*} = \frac{(21-23.2)^2}{23.2} + ... = 2.0381$

We find that $\chi^2_{table} = \chi^2_{0.05,2} = 5.991$, so we *fail to reject* and conclude that we can't find a difference in food preference in our squirrels (our $d.f. = \nu = (r-1)(k-1) = 2 \times 1 = 2$).

Finally, a few comments about contingency tables.

For 2×2 tables we can calculate something called the *Relative Risk*. This tells us what the risk is of one thing happening compared to another. For example, if we think back to the seatbelt example, what is the risk of dying if you're not wearing a seatbelt as opposed to wearing a seatbelt? We simply compare the proportions:

$$\hat{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{0.00958}{0.00124} = 7.73$$

This tells us that the risk of dying in a car crash is 7.73 times higher if you're not wearing a seatbelt (a good reason to wear your seatbelt!).

The relative risk is used extensively in medical trials and such (e.g, the risk of getting a heart attack if you take this medicine is twice that if you don't, and so on).

The relative risk is easy to calculate but does have some drawbacks. For reasons we don't want to get into, it isn't always appropriate to use the relative risk. For this reason statisticians have developed a related quantity called the *Odds ratio* which is used even more often than the relative risk. Unfortunately, it's a little too complicated to explain here. If you're interested, you can check it out on Wikipedia.

What about the assumptions? They're identical to that of the goodness of fit test:

> Random data.
>
> Smallest expected value ≥ 5.

There is a nice alternative if you violate the second assumption - it's called Fisher's exact test. Again, you can look it up on Wikipedia if you're interested.